PB-215 555

APPLICATION OF ELECTRONIC PROCESSING TO AIR QUALITY DATA

Arthur C. Stern

Robert A. Taft Sanitary Engineering Center
Cincinnati, Ohio

1957

DISTRIBUTED BY:

NTIS

National Technical Information Service
U. S. DEPARTMENT OF COMMERCE
5285 Port Royal Road, Springfield Va. 22151

PB 215 555 ✓

LIBRARY
SANITARY ENGINEERING CENTER
CINCINNATI, OHIO

APPLICATION OF ELECTRONIC PROCESSING TO AIR QUALITY DATA

Arthur C. Stern
Chief, Community Air Pollution Program
Robert A. Taft Sanitary Engineering Center
Cincinnati, Ohio

Presented at the Semi-Annual Technical Conference,
Air Pollution Control Association, Fairmont Hotel,
San Francisco, California, November 18-19, 1957.

U. S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
U S Department of Commerce
Springfield VA 22151

Application of Electronic Processing to Air Quality Data

Arthur C. Stern
Chief, Community Air Pollution Program
Robert A. Taft Sanitary Engineering Center
Cincinnati, Ohio

U. S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service

Because of the large amount of air quality data being accumulated by the National Air Sampling Network (NASN) of the United States Public Health Service, it became evident that electronic processing methods would have to be developed for data interpretation and analysis. When finally developed, it was apparent that these methods would also be useful for analysis of air quality data resulting from air sampling by state and local air pollution control agencies and by research and industrial groups. The purpose of this paper is therefore to describe the forms of computed data output and their interpretation so that other agencies can determine whether this technique would be useful to them. This paper will not attempt to go into the details of electronic digital computer programing necessary to accomplish the results described. These details have been written up in the form of a manual (1) which includes the computer program (for the IBM-650 Computer), and which is available on request.

Need for Computer

To use these methods, the agency does not need to own a computer. All it needs is an accumulation of air quality data and a conviction that the methods herein described should be used for analysis of that data. Electronic digital computer time is available on a fee basis from several

nationwide organizations which have been set up to provide such service or from local organizations now coming into existence in several large cities. Some agencies may be able to arrange for computer time gratis or at nominal charge from other public or private agencies or universities which use computers for accounting or scientific computing purposes. When local data is of such nature as to be of national interest and usefulness and is made available to the Public Health Service for inclusion in the NASN data file, they can by mutual agreement be processed as NASN data and the computed output made available for use by both the supplying agency and the Public Health Service.

Annual and Total Tabulations

From an input of up to 99,999 measurements of the same air quality variable over a period of not more than ten years, the output first tabulates for each sampling site and for each year, the variability of the air quality item under analysis by computing and printing (Fig.1):

(a) The name of the city, the number designating the site, the year, the abbreviation of the name of the item and a letter (W, S, or B) designating whether the data is for weekdays, Sundays, or both;

(b) The number of measurements of that item at that site for that year;

(c) The minimum, maximum, and average for that item at that site for that year; and

(d) The frequency distribution of all measurements of that item at that site for that year expressed as the 10, 20, 30, 40, 50, 60, 70, 80, and 90 percent decile values of the frequency

distribution -- the 50 percent value being, of course, the median value. These nine values, ~~together with the minimum and maximum values,~~ allow the distribution to be plotted graphically, preferably on special log-probability paper on which ~~a normal~~ many frequency distributions plot as a straight line. Discussion of the utilization and interpretation of such plots is beyond the scope of this paper. Suffice it to say that they are powerful tools for air quality analysis.(2,3,4)

The second tabulation is the same as the foregoing except that it covers a span of years, up to ten, rather than each year separately. The span is at the choice of the analyst who may, by choosing a succession of several year spans, obtain moving averages of several years' spread. The third tabulation is similar to the first except that it is for all sites in the city for each separate year and is differentiated in its identification by the absence of a site number. The fourth tabulation is similar to the second except that for the span of years chosen the tabulation is for all sites in the city. As in the second tabulation, the fourth allows the analyst to develop moving averages for the city as a whole. The fifth tabulation is similar to the fourth except that it is for groups of cities within a state. The sixth tabulation is for groups of states within a geographic region, while the seventh is a grand total for all sites.

Although for NASN purposes the categories used are site, city, state, region, and nation, for a state they can be set up as site, city, county, region, and state, and for a district such as the Bay Area Air Pollution Control District they can be site, city, subregion, county, and district.

Monthly and Diurnal Tabulations

In addition to these seven categories of tabulation, all of which include frequency distributions of the item under analysis, one of two other types of tabulation is also printed which includes all the information of the seven annual and total tabulations but excludes the nine deciles of the frequency distribution. The analyst may choose for this final tabulation, without any change in the basic computer program, either:

(a) distribution of the item under analysis on a monthly basis for each city ([redacted]), or

(b) distribution of the item under analysis on a diurnal basis for each city.

In the case of NASN data, which are based on 24-hour samples, monthly distribution is selected. However, where the item is measured every few hours throughout the day, diurnal distribution may be preferable. By relatively minor changes in the normal computer program tabulation by both monthly and diurnal distribution may be obtained concurrently with the seven types of annual and total tabulation and frequency distributions may be obtained on a monthly or diurnal basis.

Output Flexibility

As noted, these nine types of output will analyze up to 99,999 measurements of a single air quality item. At the analyst's option, this entire battery of outputs is available (assuming four digits of data per item) for:

(a) Up to 120 air quality measurement items, i.e., suspended particulates;

(b) Percentage of any one item to any other, i.e., percent lead in suspended particulate; or

(c) Computations involving more than two items, i.e., ratio of sodium plus potassium plus calcium to chloride.

In addition, any of these outputs may be for:

(a) All samples;

(b) Weekday or Sunday samples only;

(c) Urban samples only;

(d) Samples from any one of nine different population categories;

(e) Nonurban samples only;

(f) Samples from any one of eight different topographic categories, e.g., desert sites only; or

(g) Any noncontradictory combination of the foregoing.

Where chemical analysis is performed in more than one laboratory, only those samples analyzed in a particular laboratory may be selectively tabulated. This feature permits separate tabulation of data supplied to the Public Health Service by a state or local agency as well as merging into NASN data from all cooperating laboratories.

Because of the flexibility built into this system of data analysis, only a small fraction of the combinations of variables available to the analyst's option will ever be used. However, it is gratifying to know that the steps necessary to set up the tabulation for any possible combination requires about one minute's time.

Suppression of Output Duplication

In determining the forms of output, the most complex situation was contemplated, that of the NASN in which there would be samples from each geographic region, within the nation, from all states within each region, from a number of cities within each state, from a number of sites within some cities, and for a span of years within each site. Because

this complexity does not exist in all cases, i.e, where there is only one site in a city or only one year's data, provision has been made to anticipate situations which would lead to duplication of output and to prevent such duplication. Thus, a normal tabulation of the data for a state in which there is only one sampling site will have as annual and total outputs only those data associated with the site, since city and state outputs would duplicate site outputs in all respects except identification. However, since there are cases where it is desirable to develop certain or all classes of output tabulation ever. if they result in duplications, it is possible to suppress any or all of these procedures.

In similar manner, duplication of the listing of the city name may be suppressed where its omission improves intelligibility of the output tabulation,or retained where its use best identifies the output. Since each type of output is on a separate punched card, outputs may be grouped by category prior to automatic typing of output tabulations.

Computations

What the computer actually does with the data to develop these outputs is: first, to determine whether data is to be taken from a card; second, to select the particular data fields for analysis from among the approximately eighteen fields on a card; third, to perform the necessary computations for percentage of one data field to another, ratios between groups of data fields, etc.; fourth, to compare the result of this computation with previous results with regard to maxima and minima, to add it into sum and count registers, and to incorporate it into frequency distribution registers; and lastly, to use the sum and count registers to compute average values, and the frequency dis-

tribution registers to compute cumulative frequency distribution decile values. The rest of the computer work is "bookkeeping" to keep track of identifying information having to do with location, time, and the several variables chosen by the analyst, and to set the output up in the several forms shown.

Form of Input

The form of input necessary to achieve these results is an eighty-column punched card with the columns used as follows:

Column	Description	Limits
1	Units digit of year	0 - 9
2-3	(Month (for monthly tabulation) ((Identifying hour (for dirunal tabulation)	01 - 12
4	Region (or equivalent)	1 - 9
5-6	State (or equivalent)	01 - 98
7-8-9	City (or equivalent)	001 - 998
10	Site	1 - 8
11	Card Number	1 - 8
12	Population Class	0 - 9
13	Topographic Class	0 - 9
14-15-16-17	Sample Number	0001 - 9999
18-19	Day	01 - 31
20	Tens digit of year	0 - 9
21-28	Data	-
29-30	Laboratory Number	00 - 99
31-80	Data	-

Since it is difficult to punch this form of input card in which the month, date, and year are not in seqeuence, original data cards can be punched in simpler form as long as all digits needed in the first thirty columns are present. Conversion to the form noted may be accomplished automatically prior to computation. For merging local or state data into NASN data, the input card should preferably correspond to NASN data cards to the extent of using the same units, the same number of digits,and the same location of decimal point but not necessarily by using identical NASN card columns.

Two of the eight possible data cards have been assigned data locations to meet the needs of the NASN. These allocations are:

CARD #1

Column	Allocation	Units	Decimal Point Location
21-24	Suspended particulate	µg/CuM	XXXX.
25-28	Organic	"	XXX.X
31-33	Ca	"	XX.X
34-37	SO_4	"	XXX.X
38-40	pH	-	XX.X
41-43	Cl	µg/M3	XX.X
44-47	Ba	"	X.XXX
48-50	F	"	X.XX
51-53	Sr	"	X.XX
54-57	NO_3	"	XXX.X
58-60	Mg	"	XX.X
61-63	Na	"	XX.X
64-67	K	"	XX.X
68-70	Be	"	.XXX
71-73	As	"	X.XX
74-77	NH_3	"	XXX.X
78-80	Ag	"	.OXXX*

CARD #2

Column	Allocation	Units	Decimal Point Location
21-25	B radio-activity	µµc/CuM	XXXX.X
26-28	Mn	µg/CuM	X.XX
31-34	Suspended particulate	"	XXXX.
35-37	Sn	"	X.XX
38-40	Pb	"	XX.X
41-43	Fe	"	XX.X
44-47	Cu	"	XX.XX
48-50	Cr	"	.XXX
51-53	Ti	"	X.XX
54-57	V	"	X.XXX
58-60	Mo	"	X.XX
61-63	Cd	"	.XXX
64-67	Zn	"	XX.XX
68-70	Bi	"	.XXX
71-73	Sb	"	.XXX
74-77	Ni	"	X.XXX
78-80	Co	"	.XXX

* Data starts in the second decimal place.

For concurrent monthly and diurnal distribution analysis different data allocation would be required since the various hourly values of the item under analysis would either have to appear on the same card, or on up to eight cards of identical form.

Example of Output

An actual example of the output obtained when analyzing NASN suspended particulate data from 1953 through 1956 for New York City is shown in Figure 3. Computing time was five minutes.

Data from the last line of the annual and yearly total tabulation have been plotted (Fig. 4) to show our preferred method of presentation. All of the supplementary information on the plot is obtained from Figure 3. ~~The point of greatest interest is that the maximum actually measured in November 1953 is with 0.2 percent of highest value that could have been anticipated on the basis of this amount of data. This figure tends to give greater confidence to that portion of the curve of greatest concern, namely, the extrapolation beyond 90 percent value towards 99.99 percent.~~

These New York City data are presented to illustrate the form of NASN computer program output, not as being particularly high or low. In this use of the data comparison with other communities is not intended.

Future Work

It has taken many more months than we originally anticipated to arrive at the point where our electronic data processing methods handle large masses of data. Our next step is to apply them to data on the more than 10,000 air quality samples in our file and publish the outputs as rapidly as they are tabulated.

Because the developments of this methodology to its present state

has fully occupied our time for the last six months, no time has been devoted to the electronic computer methods for correlating air quality item analysis with weather observations. Methods for so doing are next on the agenda. Other work ahead in the field of electronic computation of air quality data is the use of the above described forms of output as input to as yet undeveloped computer programs to obtain still further statistical description and interpretation of the air quality variation.

References

(1) Stern, Arthur C. Manual on the National Air Sampling Network Program for the IBM 650 Computer. Robert A. Taft Sanitary Engineering Center, Cincinnati, Ohio.

(2) Tabor, Elbert C. and Warren, William V. Distribution of Certain Metals in the Atmosphere of Some American Cities. ~~Presented at Annual Meeting of American Industrial Hygiene Assn. St. Louis, Mo. April 25, 1957~~ (in press).

A. M. A. Archives of Industrial Health.

(3) Stern, Arthur C. Results of Three Years of Operation of the National Air Sampling Network. Presented at Golden Jubilee Meeting, Air Pollution Control Assn., St. Louis, Mo. June 5, 1957.

(4) Wanta, R. C. Application of Extreme-Value Theory to Analysis of Community Air Pollution. Transactions, American Geophysical Union, 37, 3, p. 362 (June 1956).

Acknowledgement

This development has been possible only because of the patience and cooperation of the Cincinnati IBM Applied Science personnel: Messrs. Norman Eichenberger, Donald Licking, John Ramsen, and Ronald Repking, and of R. C. Steele, Manager, Cincinnati IBM 650 Data Processing Center, all of whom were generous both with their time and the time of the IBM 650 computer. Professor Paul Herget, of the University of Cincinnati, acted as consultant to the Sanitary Engineering Center on this work. Messrs. Charles E. Zimmer, Elbert C. Tabor, and Eugene K. Harris of the Sanitary Engineering Center cooperated in the work.

Permission of the New York City Department of Air Pollution Control to use data from samples collected by their staff to illustrate this paper is appreciated.

LOCATION	S	YEAR	D	SUB	P	T	JAN	FEB	MAR	APR	MAY	JUNE	JULY	AUG	SEPT	OCT	NOV	DEC
NUMBER OF SAMPLES																		
NEW YORK N Y		5356	B	SP	0	0	15	18	26	21	17	23	21	10	23	17	18	17
MINIMUM																		
NEW YORK N Y		5356	B	SP	0	0	84	71	57	70	94	44	75	[illegible]93	81	97	79	78
AVERAGE																		
NEW YORK N Y		5356	B	SP	0	0	190	187	166	193	231	191	193	170	217	248	211	204
MAXIMUM																		
NEW YORK N Y		5356	B	SP	0	0	346	379	305	302	372	328	291	259	370	466	642	333

LOCATION	S	YEAR	D	SUB	P	T	COUNT	MIN	PER 10	20	30	40	CENT 50	FREQ 60	DIST 70	80	90	MAX	AVG
NEW YORK N Y	2	53	B	SP	0	0	3	184		198	198	198	342	342	342	342	709	642	371
	2	54	B	SP	0	0	30	145	165	198	209	220	231	245	265	285	310	309	241
	2	55	B	SP	0	0	40	93	123	149	172	198	230	258	285	331	377	456	236
	2	56	B	SP	0	0	32	120	147	173	201	209	218	227	235	273	325	364	223
	2	5356	B	SP	0	0	105	93	145	170	198	211	226	241	269	301	350	642	237
	3	53	B	SP	0	0	3	87		96	96	96	115	115	115	115	165	146	110
	3	54	B	SP	0	0	36	79	100	111	141	150	162	179	193	257	286	350	180
	3	55	B	SP	0	0	38	44	80	99	109	142	157	174	193	218	262	335	161
	3	56	B	SP	0	0	26	71	91	106	142	150	161	187	238	286	323	379	192
	3	5356	B	SP	0	0	103	44	87	106	127	147	159	178	196	246	292	379	174
	4	54	B	SP	0	0	2	162			162	162	162	162	162	165	165	164	163
	4	55	B	SP	0	0	24	44	70	80	102	127	143	[illegible]	[illegible]	42	211	238	136
	4	5455	B	SP	0	0	26	44	74	80	109	127	145	1	[illegible]82	182	198	238	138
		53	B	SP	0	0	6	87	96	96	115	115	165	198	198	342	342	642	240
		54	B	SP	0	0	68	79	111	146	157	175	198	221	246	269	306	309	207
		55	B	SP	0	0	102	44	83	106	138	153	169	190	218	259	318	466	185
		56	B	SP	0	0	58	71	115	144	160	186	206	223	238	272	323	379	209
		5356	B	SP	0	0	234	44	97	128	149	166	186	200	234	268	310	642	199

FIGURE 1

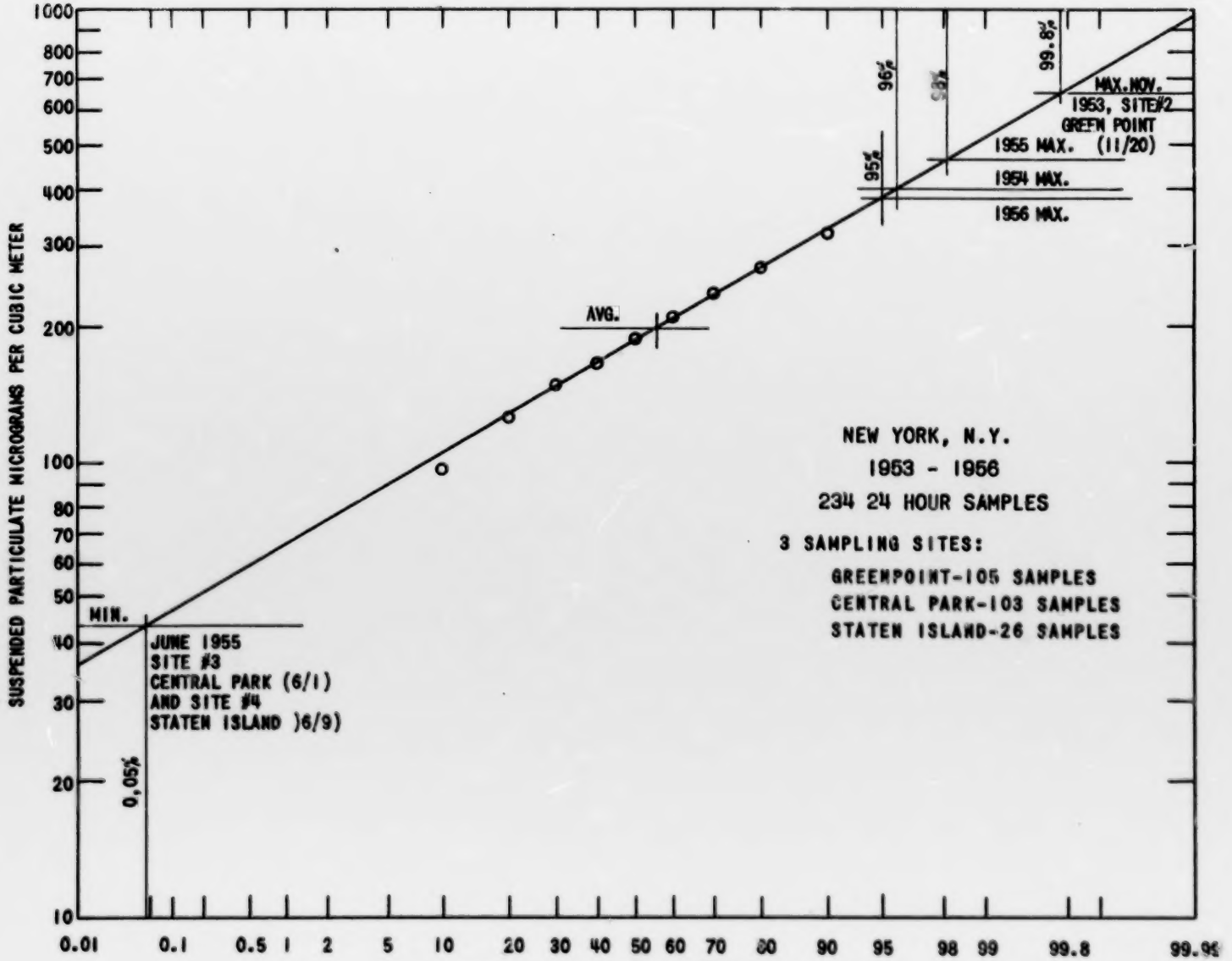
SUSPENDED PARTICULATE MICROGRAMS PER CUBIC METER
1000
800
700
600
500
400
300
200
100
80
70
60
50
40
30
20
10
0.01
0.1
0.5
1
2
5
10
20
30
40
50
60
70
80
90
95
98
99
99.8
99.95
% OF SAMPLES LESS THAN STATED MICROGRAMS PER CU/M
99.8%
96%
95%
MAX. NOV. 1953, SITE#2 GREEN POINT
1955 MAX. (11/20)
1954 MAX.
1956 MAX.
AVG.
MIN.
JUNE 1955
SITE #3
CENTRAL PARK (6/1)
AND SITE #4
STATEN ISLAND)6/9)
0.05%
NEW YORK, N.Y.
1953 - 1956
234 24 HOUR SAMPLES
3 SAMPLING SITES:
GREENPOINT-105 SAMPLES
CENTRAL PARK-103 SAMPLES
STATEN ISLAND-26 SAMPLES

FIGURE #2

41